The Industry 4.0 Case for Edge AI Accelerators

Edge AI accelerators optimize real-time processing, reduce latency, and improve energy efficiency, making them superior to traditional GPUs for edge deployments.

Current Compute Environment

In the era of Industry 4.0, AI-based computer vision enables machines to interpret visual data—essential functions for automation, quality control, and operational efficiency.

Historically, AI computer vision relied on GPUs to handle these functions. GPUs can handle heavy matrix math computations. GPUs are very good at processing graphics and scientific workloads. Additionally, training is suited for this technology.

However, GPUs are not practical for most industrial environments. They are primarily designed for desktop or data center use, making them less suitable for industrial applications. Additionally, significant expense is added wtih higher power and airflow requirements. Off-the-shelf GPU boards, such as the NVIDIA RTX 4060, do not meet industrial temperature standards. Their size and heat dissipation requirements also complicate achieving acceptable ingress protection (IP) ratings in industrial computing systems.

AI accelerators bridge the gap between traditional CPUs and the intensive tensor processing required by AI computer vision workloads. They are a core technology for realizing Industry 4.0 goals.

Benefits of Edge AI Accelerators in Industrial Environments

AI accelerators make AI processing feasible in industrial environments by delivering the necessary performance in a smaller footprint with lower power dissipation. Using AI accelerators reduces system costs, boosts reliability, and lowers total ownership cost (TCO).

Memryx AI Accelerator

The MemryX AI Accelerator supports many AI models originally developed for GPUs. Thousands of GPU-targeted models can be compiled to run on the MemryX AI Accelerator without modifying the model source code. This compatibility provides a straightforward way to replace a GPU with a MemryX AI Accelerator in typical x86 or ARM-based industrial computing systems.

A steroscopic camera for 3D imaging measuring time of flight depth sensor in a factory.

Key Use Cases

Quality Control and Inspection

AI computer vision systems automate product inspections directly on production lines, identifying defects with high precision and consistency.

These systems improve product quality, reduce human error, and operate faster than manual checks. In manufacturing, computer vision spots flaws in components within milliseconds on high-speed lines. Quick identification helps maintain high standards and minimizes waste.

Autonomous Navigation and Vehicle Guidance

Autonomous Guided Vehicles (AGVs) use AI vision to transport materials without human intervention. Factories and warehouses benefit from labor cost savings and streamlined logistics by tracking materials at rest and in transit.

Reports suggest using robots in warehouses can reduce human walking by up to 80%.

Worker Safety and Compliance

AI computer vision systems monitor worker behavior to ensure compliance with safety protocols. These systems alert supervisors when workers lack protective equipment or enter restricted areas.

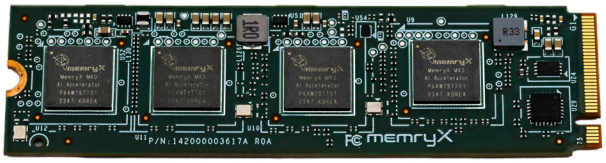
In high-risk environments like construction or areas near robots, rapid detection of non-compliance helps maintain safety standards and lowers accident rates.

Process Optimization

AI-powered computer vision revolutionizes process optimization in logistics by providing real-time insights. It detects conveyor jams, misrouted packages, or equipment malfunctions in logistics hubs, prompting immediate corrective action.

With faster, more accurate visual processing, AI vision enables predictive insights and operational improvements. AI vision is vital for modern logistics and operational management.

The Role of MemryX MX3



AI accelerators are purpose-built to address the specific AI computational needs of edge environments like factories.

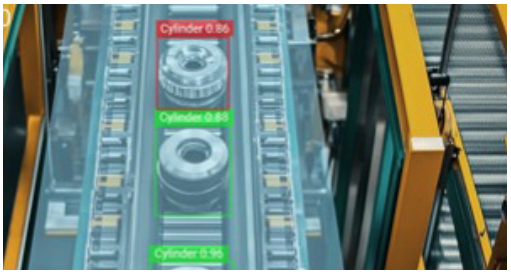
Unlike general-purpose processors, the MemryX MX3 is optimized for edge AI inference with a focus on power efficiency, low latency, and high throughput. This approach is ideal for real-time industrial applications where energy efficiency and fast processing are crucial for operational success.

The MX3 provides excellent AI computer vision performance and accuracy for applications such as autonomous navigation and automated quality inspection. It can be added to industrial computing systems (IPC) via M.2 or PCI Express slots and supports x86 or Arm processors under Windows or Linux.

The MemryX MX3 is the best way to add AI processing power to industrial systems without compromising size, weight, power (SWP), ruggedness, and reliability. The MemryX MX3 is the key to achieving computer vision success in Industry 4.0.

AI Accelerator Capabilities

Defect Detection



On high-speed production lines, the MX3 accelerates AI models used for defect detection and enables real-time analysis. Models are trained to identify specific defects and can be updated in the field as new defect types emerge. This adaptability and longevity are valuable in dynamic production settings.

Common models for defect detection include ResNet, VGG, and YOLO, all of which run successfully on the MX3 platform. These AI vision methods outperform prior approaches in speed and accuracy, especially in complex scenarios.

Semantic Segmentation

Semantic segmentation models label each pixel in an image to identify various elements. This output is often visualized as color overlays. Segmentation assists path planning software in identifying roads or other areas to navigate or avoid. The YOLO model family offers segmentation-optimized versions.

The MX3 demonstrates high performance with YOLO, DeepLabV3, MediaPipe, and Unet. Many of these models run three times faster on the MX3 than on GPUs.

Object Detention & Pose Recognition

The MX3 quickly identifies safety issues such as missing Personal Protection Equipment (PPE). Models trained to detect proper PPE usage generate alerts when they find unsafe environments, increasing workplace safety.

Pose recognition combined with object detection also improves factory safety by identifying individuals in distress or in hazardous positions. Models like MediaPipe Holistic compile efficiently on the MX3, enabling these critical applications.

MemryX MX3 AI Module vs. GPU in Edge Computing

GPUs have been central to advancing AI by supporting parallel processing for algorithms such as convolutional neural networks (CNNs). They helped overcome significant barriers, enabling computers to sense in new ways.

Despite their importance in AI's growth, GPUs are unsuitable for many industrial environments. Better solutions are required. MemryX enables developers to compile GPU-based AI models and run them on more efficient accelerator architectures tailored for industrial use.

The MemryX MX3 is designed to provide efficient, real-time edge processing with minimal energy consumption. This suits power-sensitive Industry 4.0 applications, including autonomous navigation, quality inspection, and safety monitoring. The MX3 chip is very small—less than one square centimeter—and does not need external memory. Multiple chips can be cascaded for low-cost or high-performance needs. MemryX offers an M.2 module with four MX3 chips that can deliver up to 26 TFLOPS. Systems of various sizes can be built using MX3 as a foundational AI processing building block.

The MX3 reduces system size compared to GPUs due to the at-memory architecture and compact size. For example, the MX3 M.2 AI Accelerator only occupies about 3% of the volume and uses less than 10% of the power of an RTX 4060 GPU with comparable performance. The MemryX MX3 with four accelerators costs less than half the price of most GPUs.

Comparison

	MemryX MX3 M.2 AI Module	NVIDIA RTX 4060 GPU	Bottom Line
Performance	380 frames/sec	328 frames/sec	Even one FPS faster means better models and analysis
Power	10 Watt TDP 5 Watt Typical	120 Watt TDP 110 Watt Typical	Enables edge compute in almost any location. Solar can be used vs. full power.
Size	22 x 80 x 10 mm	240 x 111 x 40 mm	Less footprint; smaller PCs can be used in places where space is at a premium.

94% In Deloitte's "2022 Global AI in Manufacturing Survey," 94% of manufacturing executives indicated that AI was key to their growth strategy, with computer vision specifically cited as a technology being integrated for visual quality inspection and predictive maintenance.

MX3 Value Propositions

The MX3 offers high throughput and accuracy for industrial AI vision models.

- **Performance**: Processes up to 350 frames per second with YOLOv8, maintaining per-frame latency under 10ms. Production speeds remain unaffected.
- **Accuracy**: Benchmark tests demonstrate detection accuracy above 95% for defects such as scratches, dents, and discoloration, based on precision and recall measurements.
- **Efficiency**: With an optimized architecture, the MX3 achieves up to 60 Frames/Sec/Watt, making it ideal for edge deployment.

Conclusion

A primary goal of Industry 4.0 is to boost productivity by making production processes more efficient. AI accelerators help achieve this by integrating intelligence directly into the workflow. Industry 4.0 also aims to improve flexibility, allowing production systems to adapt quickly and economically to changing needs, including personalized manufacturing. AI accelerators play a significant role in enabling this adaptability via system connectivity and programmability.

MemryX MX3's innovative at-memory architecture makes AI both feasible and cost-effective in factory environments by resolving the limitations of traditional GPUs. The MX3 is in production today. To explore how the MemryX MX3 can enhance your Industry 4.0 plans or to discuss deployment options, please contact us at memryx.com/contact.

1600 Huron Parkway, 2nd Floor, Ann Arbor, MI 48109, USA

MemryX Inc. is an AI semiconductor startup headquartered in Ann Arbor, MI USA with branches in Taipei and Hsinchu, Taiwan. We develop a highly scalable and innovational AI accelerator offering high performance, low power, and customer ease of implementation for embedded Edge AI vision-based applicationas and real-time processing.